SLC Walkthrough for Practice Set Week4

- Import Required libraries.
- Load the dataset.

Let's begin with some hands-on practice exercises:

1. Is there any record where no data have been reported? If yes, do the needful.

HINT:

- Check, If any missing values are presented or not.
- If you find any missing values, Handle them by removing the rows or columns or fill with mean or median or mode.

2. Use the visualization technique to identify the variables with missing data.

HINT:

Use heatmap to visualize the missing values.

3. Use the appropriate technique to replace the missing data in the variable representing the length of a wheat kernel.

HINT:

- Check the skewness or distribution of the variables which are having missing values.
- If there nearly normal distributed fill with mean or else fill with median.
- If missing values present in the categorical variable fill with the mode.

4. Identify the variables containing extreme values and remove such observations, if present.

HINT:

- Check the outliers using related distribution plots.
- If you find any extreme values, Use IQR method to remove them.

5. Apply the gradient boosting on 80% of the data with 250 estimators each of maximum depth 2. Also, calculate the precision and sensitivity using the confusion matrix.

HINT:

- First split the data into train and test data.
- Build a gradient boosting model on a training dataset with the given parameters.
- Calculate the Precision and Sensitivity using confusion matrix.

6. Create 80 stumps using AdaBoost and plot the ROC curve along with the AUC score.

HINT:

- Use the train and test set from Q5.
- Build an Adaboost model on a training dataset with number of estimators are 80.
- Plot the ROC curve along with the AUC score.

7. Select the optimal maximum depth from the given values for 180 base learners to build the gradient boosting classifier (consider 3-fold cross validation).

Use the given list:

depth = [2, 3, 4, 5, 6, 7, 8]

HINT:

- Use Grid Search CV to tune the params, use 3 fold cross validation to tune the hyper parameters in Grid Search CV.
- Get the best param from the grid model.

8. Build the XGBoost model with a learning rate of 0.4 and gamma equal to 3. Calculate the accuracy by plotting the confusion matrix.

HINT:

- Build an XGBoost model on a training dataset with learning rate is 0.4 and gamma is 3.
  Consider the train and test set in Q5.
- Plot the confusion matrix and calculate the accuracy of the model.

9. Use the stacking technique on 70% of the data with the 9-NN and Naive Bayes model as base learners. Consider the Adaboost model as a final estimator. Also, compute the AUC score of the model.

HINT:

- Scaling is must for KNN classifier, Thus scale the features.
- Split the scaled data into train and test sets.
- Build the stacking classifier using the KNN (with K = 9) and Naive Bayes model as base learners. Consider the Adaboost classifier as a final estimator.
  Ex: base_learners = [('model name', model_classifier(parameter1 = x)),
  ('model name, model_classifier())]

  StackingClassifier(estimators = base_learners, final_estimator = final_classifier(parameters if any))

- Calculate the AUC score of the model.

10. Use the base learners in Q9 and build a stacking model with the XGBoost as final estimator. Compute the AUC score of the model and compare it with the result of Q9.

HINT:

- Follow the similar steps of question 9, only change final model as xgboost.
- Compare the AUC score with previous model which you built in question 9.